Big Data Processing and Analysis Course

Movie Rating Prediction With User Based CF

Mustafa Katipoğlu

16011084

General Info about Project:

Our application aim to provide personalized movie recommendations to the users. In order to provide recommendations, we have used user-based collaborative filtering.

We have implemented 5 main functions in Hadoop:

- Movie Rating Prediction
- Pearson Correlation in between users
- K Nearest Neighbor For Each User
- Average Rating Per User
- Average Rating Per Movie

Technical Challenges:

One of the biggest challenges in the development process was to implement Movie Rating Prediction, Pearson Correlation and K Nearest Neighbor.

- With K Nearest Neighbor, we needed to sort the data and partition the data in a different way other than the one Hadoop provides by default. That is why we had to implement our own group comparator, partitioner and our own custom Writable data types for sorting the data.
- When calculating the Pearson correlation and movie prediction, the default Hadoop data types was not good enough for us. We have implemented our own Writable data types here as well.
- And also when calculating these 3 functions we have to pipeline results so that result of one becomes input of another.
- In some cases, we had to get data from multiple data sources, join them by using map-join methodology.

Implementation Details

Our main aim was to make movie rating prediction that is why almost all other functionalities is part of this function.

We first start by identifying average user ratings, and common movies rated by users in order to implement Pearson correlation.

Then we sort the Pearson correlations for each user by implementing our own secondary sort, and then get the K biggest of them for each user.

Last but not least, we get the K nearest neighbor and their movie rating for the movie we want to make prediction on. By taking weighted average of Pearson correlation of each neighbor and their rating to the movie, we calculate the prediction rating for the user of interest.

Performance Evaluation

As an extra to Hadoop implementation we have implemented same functionalities in normal python by using pandas and numpy. We compare the results to each other in order to detect anomalies.

Experience

As we try to develop movie prediction in Hadoop, we have gained a lot of knowledge about inner working of Hadoop as well as got practice in different parts of it.

Here is the list of topics we have actively used and explored as part of the project:

- Creating Custom Hadoop Writable Datatypes
- Secondary Sort
- Chain Mapper, Chain Reducer
- Map Join – Reduce Join